Electronic healthcare databases in Europe

Pacurariu, Alexandra; Plueschke, Kelly ; McGettigan, Patricia ; Morales, Daniel; Slattery, Jim ; Vogl, Dagmar

Published in:
BMJ Open

DOI:
10.1136/bmjopen-2018-023090

Publication date:
2018

Licence:
CC BY-NC

Document Version
Publisher's PDF, also known as Version of record

BMJ Open

Electronic healthcare databases in Europe: descriptive analysis of characteristics and potential for use in medicines regulation

Alexandra Pacurariu,[1] Kelly Plueschke,[1] Patricia McGettigan,[1,2] Daniel R Morales,[1,3] Jim Slattery,[1] Dagmar Vogl,[1] Thomas Goedecke,[1] Xavier Kurz,[1] Alison Cave[1]

To cite: Pacurariu A, Plueschke K, McGettigan P, *et al*. Electronic healthcare databases in Europe: descriptive analysis of characteristics and potential for use in medicines regulation. *BMJ Open* 2018;**8**:e023090. doi:10.1136/bmjopen-2018-023090

► Prepublication history and additional material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2018-023090).

Received 20 March 2018
Revised 31 July 2018
Accepted 2 August 2018

[1]Department of Surveillance and Epidemiology Service, European Medicines Agency, London, UK
[2]William Harvey Research Institute, Queen Mary University of London, London, UK
[3]Division of Population Health Sciences, University of Dundee, Dundee, UK

Correspondence to
Alexandra Pacurariu;
alexandra.pacurariu@ema.europa.eu

ABSTRACT

Objective Electronic healthcare databases (EHDs) are useful tools for drug development and safety evaluation but their heterogeneity of structure, validity and access across Europe complicates the conduct of multidatabase studies. In this paper, we provide insight into available EHDs to support regulatory decisions on medicines.

Methods EHDs were identified from publicly available information from the European Network of Centres for Pharmacoepidemiology and Pharmacovigilance resources database, textbooks and web-based searches. Databases were selected using criteria related to accessibility, longitudinal dimension, recording of exposure and outcomes, and generalisability. Extracted information was verified with the database owners.

Results A total of 34 EHDs were selected after applying key criteria relevant for regulatory purposes. The most represented regions were Northern, Central and Western Europe. The most frequent types of data source were electronic medical records (44.1%) and record linkage systems (29.4%). The median number of patients registered in the 34 data sources was 5 million (range 0.07–15 million) while the median time covered by a database was 18.5 years. Paediatric patients were included in 32 databases (94%). Completeness of information on drug exposure was variable. Published validation studies were found for only 17 databases (50%). Some level of access exists for 25 databases (73.5%), and 23 databases (67.6%) can be linked through a personal identification number to other databases with parent–child linkage possible in 7 (21%) databases. Eight databases (23.5%) were already transformed or were in the process of being transformed into a common data model that could facilitate multidatabase studies.

Conclusion A Few European databases meet minimal regulatory requirements and are readily available to be used in a regulatory context. Accessibility and validity information of the included information needs to be improved. This study confirmed the fragmentation, heterogeneity and lack of transparency existing in many European EHDs.

Strengths and limitations of this study

- Data extraction was based on information provided by database owners and publicly available information.
- Incomplete data extraction cannot be excluded, especially for very small databases with few published outputs.
- Validation of the data source was evaluated indirectly through the validation studies reported by the database owners.
- The inventory was endorsed by an expert working group of the European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (the ENCePP Working Group 'Data Sources').

INTRODUCTION

The European Union (EU) medicines regulatory network has responsibility for protecting patients by ensuring continuous evaluation of the safety of authorised medicines. At the core of such review is the scientific assessment of all available evidence including relevant information from the literature, results from non-clinical studies, randomised clinical trials, observational studies, spontaneous reports and results of other available research. A way to collect more information about a medicine's safety postmarketing is by means of postauthorisation safety studies (PASS).[1] PASS may be imposed on a marketing authorisation holder by a regulatory authority or conducted by the company to address a safety concern or evaluate the effectiveness of risk-minimisation measures aimed at reducing the occurrence or severity of an adverse reaction.[2 3]

Secondary use of routinely collected data from electronic healthcare databases (EHDs) is often used in such studies because it is usually faster and cheaper than primary data collection.

A review of pharmaceutical industry-sponsored studies evaluating the effectiveness of risk-minimisation measures submitted to the European Medicines Agency (EMA) for

cardiovascular, endocrinology or metabolic drugs authorised between 1995 and 2015 found that EHDs were used in 53% of studies evaluating routine risk-minimisation measures and in 31% of studies evaluating additional risk-minimisation measures.[4] A second review of 189 PASS assessed by the EMA between 2012 and 2015 and registered in the EU electronic Register of Post-Authorisation Studies (EU PAS Register) reported that secondary use of routinely collected data was found in 33.3% of PASS, and 58% among these leveraged electronic health records (EHRs).[5] A third review of a different set of studies registered in the EU PAS Register as of December 2016 found that 117 studies (37%) used an existing claims or electronic medical records database.[6] A fourth review evaluating studies which measured the impact of regulatory interventions found that claims databases were used in 45% of studies, while EHRs were used in 22% of them, the latter being the most used type of data sources for such studies.[7] The frequent use of EHDs in observational studies was also reported in a wider context in a review of the abstracts of presentations made at the International Conference for Pharmacoepidemiology: 53% (in 2000) and 51% (in 2005) of submitted EU pharmacoepidemiological studies were conducted using automated general practice, pharmacy or claims data.[8]

The fact that between 30% and 50% of observational postauthorisation studies use EHDs as their main data source reflects the importance of these data sources to support regulatory decision-making.[1 9] On the other hand, the use of EHDs in preauthorisation research is currently limited and mostly focused on providing historical control data or understanding the natural history of the disease.

As regulatory decisions based on EHDs may have a considerable impact on public health, the quality of the information, the validity and reproducibility of the derived results require close attention, especially when combining data from several data sources or when the original data is transformed before analysis.[10–12] It has been emphasised that the same level of scientific rigour should be employed irrespective of the study design and data source to be used, and that the strengths and weaknesses of each data source should be considered.[13] The speed at which the results could be generated is an additional important consideration, particularly for regulatory purpose.[9 14 15] By considering the characteristics of the data sources and the research objectives to be addressed, the investigators should be able to choose the most appropriate resource(s) to address the question at hand. However, while some authors provide a detailed description of the databases used in their study,[16–19] in other cases the description is often incomplete, and a justification for their choice in the context of alternative data sources is rarely provided.[18] The International Society of Pharmacoepidemiology has developed guidelines to support the selection and use of data sources for observational research by highlighting potential limitations of databases and recommending testing procedures. The guidelines also provide a checklist covering six areas: database selection, use of multiple data resources, extraction and analysis of the study population, privacy and security, quality and validation procedures and documentation.[20] The availability of an inventory of European databases describing the main characteristics, conditions of access and validation performed would support investigators to identify databases suitable for their research question. Moreover, knowledge of the characteristics of the data sources used in a postauthorisation study would enhance regulators' confidence in the evidence derived from such data and ultimately in the usefulness of the study in the decision-making process.[21 22]

The main objective of this study is to provide an inventory of EHDs and describe their key characteristics and availability with the aim to support stakeholders in their choice of the data source when conducting a postauthorisation study.

METHODS

Identification of EHDs

As a first step, we identified existing EHDs in Europe by screening the following sources: the European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP) resources database,[21] web-based search engines,[22] textbooks on clinical pharmacoepidemiology,[23 24] publicly available inventories created for European Commission-funded research projects and databases used in EMA-funded postauthorisation studies.

As a second step, data sources were included in the inventory based on the following regulatory relevant criteria: the data are available to regulatory authorities or to third parties for research purposes; the database contains information on both drug exposure and health outcomes and is not disease or product specific; there is longitudinal data capture. Provision of relevant data for benefit–risk decision-making was one of the key criteria for selecting studies meeting regulatory requirements.

Prescription-only databases were excluded because they cannot be used for aetiological studies in the absence of the outcome recording. Product-specific or disease-specific registries were considered out of scope as they create cohorts of patients whose entry is defined either by exposure to a product or by occurrence of a disease or health outcome.[25]

Product-specific registries are frequently used for the benefit–risk monitoring of specific products, however they rarely cover a wide range of medicines and health conditions and have a narrow scope. Databases where the data collection ceased and the historical data were not accessible were also excluded.

Data extraction and classification

For each database, publicly available information (on the databases' websites or in publications) was supplemented by contacting data source owners in writing. A total of 82% database owners responded. Teleconferences with

seven database owners were conducted to clarify some of the information provided.

The information was extracted by six EMA reviewers (AP, KP, PMG, DM, JS, AC), and the entries for each database were cross-checked for consistency by a second reviewer. Uncertainties about the classification of any variable were resolved through discussion.

The data sources were classified in three categories according to their structure, purpose and type of data: electronic medical records, claims databases and healthcare record linkage systems (eg, several databases are linked to form a complete database).

The different population registries within the same country were considered as a single national EHD if they could (and are routinely used as such) be linked using a unique identification number (eg, in Nordic countries and in Scotland).The size of the data source was quantified by the cumulative number of patients included (both total and active patients) and number of years since the initiation of data collection in the database.

Data collected in the following categories was also recorded: demographic information (age and gender of each individual), information on prescribed or dispensed medicines (including name, dose, duration, route of administration and therapeutic indication), immunisations, diagnosis data and referrals for laboratory investigations, imaging and other procedures. Information of laboratory tests results was not collected.

Availability of validation studies

Database owners were asked to report validation studies which they were aware of for their database. Studies published up to September 2016 were included. For the purpose of this study, a validation study was defined as any study published in a peer-reviewed journal that aimed to validate the information available on an outcome or exposure in comparison with gold standard information, usually the patients' original health records as reviewed by a medical professional or the same information captured by another database for a different purpose. For example, a study in 2012 compared cancer records in a general practitioners' database, hospital records and cancer registries and found considerable discrepancies in cancer recording between these different data sources.[26]

Accessibility

The accessibility of databases for research purposes was classified in four categories: no access, indirect access through the database owner or a third party, direct access restricted to specific datasets and direct access to the full dataset.

Coding of database characteristics for usefulness for medicines benefit–risk evaluation

Instead of evaluating the quality of each database, we aimed to assist in the selection of databases by implementing a coding process that identifies the data sources considered to provide sufficient information to contribute to regulatory questions on the benefit–risk evaluation of medicines. For this purpose, the following domains were included in the coding process: extent of data capture of study variables, size of data source, quality and validity of information, accessibility, potential for linkage and existing process in place to convert the data into a common data model (CDM) (figure 1). A CDM provides a common representation and architecture of the data across multiple databases, thus enabling the standardisation of administrative and clinical information and allowing the use of common analytical tools.[27]

The ENCePP Working Group 'Data Sources'[9] reviewed an initial version of the inventory with the description of databases and endorsed the final inventory.

Patient involvement statement

This descriptive analysis did not involve any patients.

RESULTS

General overview

The initial search generated a list of 77 potential data sources. After merging the national registries into a single entry and applying the exclusion criteria, 34 of them were retained in the final inventory (figure 2). Table 1 provides a list of these 34 databases and the complete information is provided in the online supplementary material.

The most frequent types of data source identified were electronic medical records (n=15, 44.1%) followed by record linkage systems (n=10, 29.4%) and claims databases (n=9, 26.5%). In terms of the type of care covered, mixed-care settings (primary and secondary care) were most common (n=17, 50%), followed by primary care databases (n=11, 32.3%) (table 2). The median number of patients followed cumulatively across the 34 data sources was 5 million (range 0.07–15 million).

Patient age and gender were recorded in all data sources while paediatric patients were included in 32 databases (94%). The median year for database start was 1998, with the oldest database established in 1964 (the Finnish Hospital Discharge Register). The median calendar time covered by a database was 18.5 years (range 7–53 years). In terms of geographical coverage, 17% of databases collect data from Norway, 14% from Finland and 10% from Denmark and Italy (figure 3).

Information captured

By definition, all the databases retained in the final inventory contained information about drug exposure (either prescribed or dispensed). The completeness of information was however variable: 28 databases (82.3%) had information about prescribed dose and duration of treatment (either directly recorded or inferred from other collected variables); 14 (41.1%) had information about route of administration; 20 databases (58.8%) recorded the therapeutic indication associated with the prescription (either directly recorded or inferred from other database elements). Over-the-counter drugs were rarely

Figure 1 Coding of the characteristics of electronic healthcare databases available in Europe for the benefit–risk evaluation of medicines. The coding system was binary: 0 if information was absent and 1 if it was present. The degree of completion for a specific variable was not recorded. An exception to the binary classification was done for the accessibility variable: 0, no access; 1, indirect access through database owner or third party; 2, direct access to specific data sources; 3, direct access to full data source.

and inconsistently captured in any of the databases while vaccinations were captured in 13 databases (38.0%). Data on hospital inpatient administered drugs were rarely captured (5.8%).

All databases had information about medical events (diagnosis) as a prerequisite for inclusion in our inventory. Referrals for laboratory investigations were captured in 19 (55.9%) and referrals for imaging or other diagnostic procedures were captured in 16 databases (47.1%).

Validation studies

No published validation study was reported for 17 databases (50.0%), while a total of 42 validation studies were reported for the other 17 databases, with a median of 3 validation studies per database (range: 1–25). The validation concerned either specific health outcomes or prescription information. The most common gold standards used for the validation included paper-based prescriptions, medical records, death records and perinatal deaths obtained from registries or national statistics reports. Some database owners have reported as validation studies the validation of prediction algorithms for various health outcomes as chronic kidney disease, ischaemic stroke and various types of cancers based on an estimating the absolute risk of a particular outcome in primary care patients with and without symptoms.[1 2] It is debatable if these are truly validation studies according to our definition.

Accessibility and potential for linkage

During selection, one database was excluded due to lack of access to third parties. From the 34 included databases, 11 (32.3.4%) offer indirect access to the database for third parties, 6 (17.6%) provide direct access to specific datasets and 11 (32.3%) offer direct access to the full dataset. The level of access could not be identified for 6 EHDs (17.6%). In terms of linkage, 23 databases (67.6%) could be linked through a unique personal identification number (PIN) to other databases containing additional healthcare-related information including cause of death registries, hospital data, prescription databases and cancer registries. The Nordic

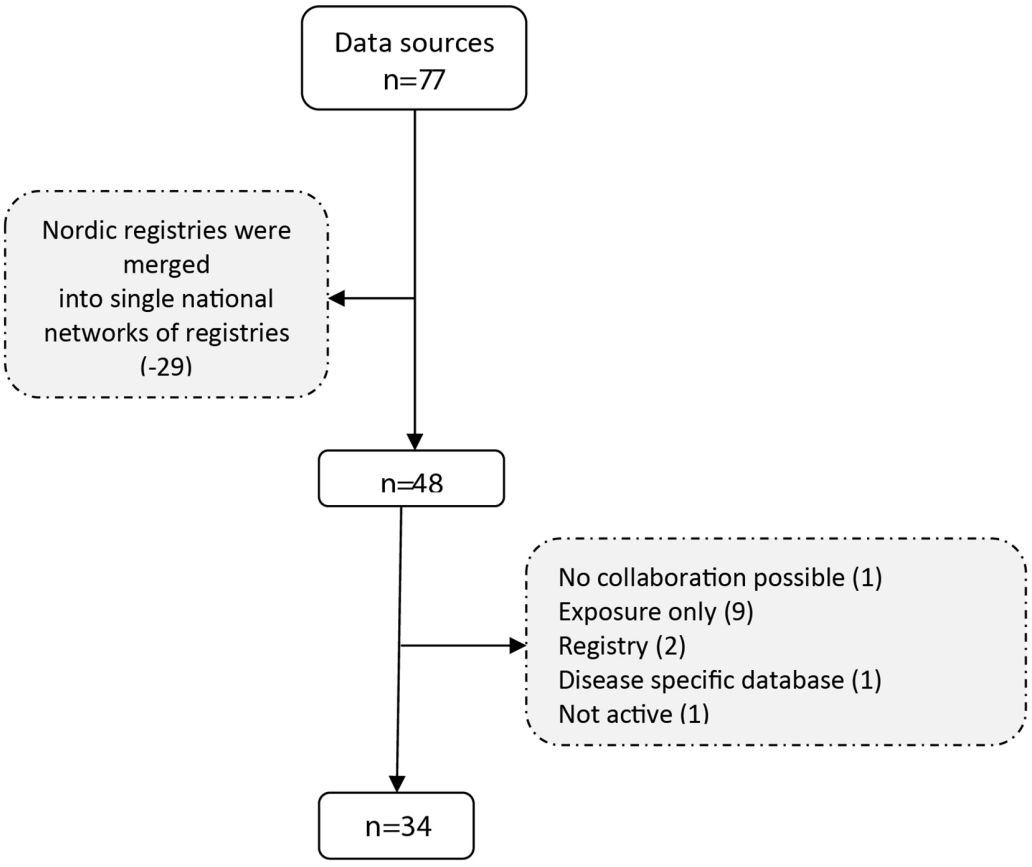

Figure 2 Flow chart of database selection.

registries are a good example of extensive linkage among different national registries through usage of a PIN. Other forms of linkage do exist, for example, in order to avoid the use of PIN and preserve anonymity, the PHARMO network uses probabilistic linkage based on patient birth date, gender and general practitioner code. The linkage of a parent with their child ('parent–child linkage'), which is useful for studies investigating pregnancy exposures and effect on offspring, was available in seven data sources (20.6%).

Conversion of the database to a CDM

Four (11.7%) databases were already transformed in a CDM and four others were in the process of being converted to a CDM (the QuintilesIMS Disease Analyser France and Germany, the Spanish Information System for the Development of Research in Primary Care, the Agenzia Regionale di Sanità Tuscany database, The Pedianet, the Clinical Practice Research Datalink, the Integrated Primary Care Information Database and The Health Improvement Network). Seven of these eight databases used the Observational Medical Outcomes Partnership CDM,[27] while the Spanish Information System for the Development of Research in Primary Care[28] is implementing the model used in the Accelerated development of vaccine benefit-risk collaboration in Europe (ADVANCE) project for vaccine studies.[29]

DISCUSSION

A total of 34 European EHDs with potential for use in the regulatory environment were included in this study. The most frequently represented regions were Northern, Central and Western Europe, with a scarcity of data sources in Eastern Europe. The most common data sources assessed were electronic medical records with a mix of primary and secondary care coverage. Most of the databases contain outpatient prescribing while inpatient prescribing is very rarely captured. The median number of patients registered within the 34 data sources was 5 million, and the median calendar time covered by a database was 18.5 years. In terms of accessibility, 24% of databases offered direct access to the full data source, with the rest having a somewhat more limited access. There are a few similar studies of EHDs available in Europe,[8 30] but as far as we are aware this is the first study taking a regulatory perspective. An analysis of the characteristics of postauthorisation studies requested by regulators showed that 47% of studies involved secondary use of data emphasising the important role of secondary data in the regulatory setting. More detailed descriptions of database characteristics are provided in electronic repositories such as the European Medical Information Network (EMIF), the ENCePP resource database and the Bridge to Data initiative.[21 22] However, existing repositories are either incomplete, have a limited coverage or

Table 1 List of data sources retained in the final inventory (by year)

Data source name	Country	Type	Type of care	Start date
Finnish National Registries	Finland	Record linkage system	Mixed	1964
Swedish National Registries	Sweden	Record linkage system	Mixed	1970
Danish National and Regional Registries	Denmark	Record linkage system	Mixed	1977
The electronic Data Research and Innovation Service	Scotland	Record linkage system	Mixed	1981
Clinical Practice Research Datalink	UK	Electronic medical records	Primary care	1987
QResearch	UK	Electronic medical records	Primary care	1989
Information System of Parc de Salut del Mar	Spain	Electronic medical records	Primary care	1990
Medicines Monitoring Unit Scotland	Scotland	Record linkage system	Mixed	1990
PHARMO Database Network	Netherlands	Record linkage system	Mixed	1990
QuintilesIMS Disease Analyser	Germany	Electronic medical records	Mixed	1992
Integrated Primary Care Information Database	Netherlands	Electronic medical records	Primary care	1995
Agencia Regionale di Sanita Tuscany database	Italy	Claims	Secondary care	1996
Norwegian Registries	Norway	Record linkage system	Mixed	1997
QuintilesIMS Disease Analyser	France	Electronic medical records	Primary care	1997
Region Emilia Romagna Database	Italy	Claims	Secondary care	1997
Hospital Information System—Lazio	Italy	Claims	Secondary care	1998
Icelandic Registries	Iceland	Record linkage system	Mixed	1998
Pedianet Database	Italy	Electronic medical records	Primary care	1998
Securite Sociale de l'Assurance Maladie	France	Claims	Mixed	1999
Lombardia Health Database	Italy	Electronic medical records	Secondary care	2000
QuintilesIMS Longitudinal Patient Database (LPD)	Italy	Electronic medical records	Primary care	2000
QuintilesIMS LifeLink: Hospital Disease Database—Belgium	Belgium	Electronic medical records	Secondary care	2001
QuintilesIMS LPD Health Search Database	France	Electronic medical records	Mixed	2001
Base de Datos para la Investigación Farmacoepidemiológica en Atención Primaria	Spain	Electronic medical records	Primary care	2002
Caserta Database	Italy	Claims	Primary care	2002
The Health Improvement Network	UK	Electronic medical records	Primary care	2002
German Pharmacoepidemiological Research Database	Germany	Claims	Mixed	2004
Echantillon Généraliste de Bénéficiaires	France	Claims	Mixed	2006
QuintilesIMS LPD Health Search Database Longitudinal	Spain	Electronic medical records	Mixed	2006
The Information System for the Development of Research in Primary Care	Spain	Electronic medical records	Primary care	2006
VEKTIS	Netherlands	Claims	Mixed	2006

Continued

Table 1 Continued

Data source name	Country	Type	Type of care	Start date
Secure Anonymised Information Linkage	Wales	Record linkage system	Mixed	2007
National Health Fund	Poland	Claims	Mixed	2008
Hospital Treatment Insights	UK	Record linkage system	Secondary care	2010

they require a fee for access, therefore restricting access to their information.

This study helps identify databases with key characteristics as an entry door to further investigate with their owner their potential usefulness for a specific study.

Given that different national guidelines and clinical practice can generate significant heterogeneity in how healthcare is delivered and recorded,[31] it is important that regulators have access to data from as broad a geographical spread as possible. Thus, there is a clear need for the development of data sources in EU member states which currently either have no data sources or are poorly represented.

The data recorded in the databases include some limitations. First, the limited capture of inpatient prescribing poses a problem for regulators and investigators since many newly approved drugs are specialised drugs, used exclusively in secondary care.[32] Second, some disease-specific variables (eg, biomarkers, laboratory tests and genetic data) are only exceptionally recorded, and they are required more and more often in study protocols. High-quality disease registries can to some extent meet this need in specific disease areas but they rarely capture comedications, comorbidities and adverse reactions. Improvements in the quality of inpatient care and in the recording of laboratory tests would be of value for epidemiological investigations on determinants for health outcomes, including drug-related safety issues.

With regard to validation, 50% of databases had at least one validation study published. Validation should normally be done for the data elements collected in every study. Publication of validation studies is not an indicator of the overall validity of the database but may inform researchers on the feasibility to perform study-specific validation in a database. A repository of validated outcomes in specific databases would reduce duplication of work. Such a repository should include a clear description of the methodology and limitations of the analysis.

Table 2 Distribution of data sources type and type of care covered

Type of data source	Primary care, %	Secondary care, %	Mixed, %
Claims	1 (2.9)	3 (8.82)	5 (14.7)
Electronic medical records	10 (29.4)	2 (5.9)	3 (8.8)
Record linkage system	0 (0)	1 (2.9)	9 (26.5)

Extending approved adult indications to the paediatric population is increasing and according to the European Commission's report between 50% and 90% of the medicines currently used in paediatrics have neither been tested on nor authorised for use in children.[32] Availability of real-world data is therefore particularly important for this purpose. In our review, we found that 94% of databases have some information about paediatric patients but no in-depth analysis of the available information was undertaken. A more detailed review of paediatric databases was undertaken by Neubert *et al* who concluded that in Europe, drug utilisation and outcome data are available for ~4 million children.[33] However, similar to our study, the authors highlight that efforts should be made to increase availability of inpatient data, a setting where the greatest prescribing of novel medicines occurs.[33]

While validity studies were published for half of the databases, van Staa and Klungel[15] highlighted that systematic measurement of data quality is lacking in most databases. As such and in line with the recommendations of Hall *et al*,[20] we encourage data holders to document the basic characteristics of their data source and to highlight when a change in recording practices occurs.

A new way forward to increase the speed and power of multicentres studies is the use of a CDM.[34] The advantage of using a CDM is that the transformed databases can be more easily integrated for research across a network. Although less than one-third of databases were already converted or in the process of being converted to a CDM in Europe, these figures are likely to change fast due to ongoing initiatives such as EMIF[35] and the European Health Data Network project.[36]

Access to databases for research purposes can be provided at patient level in only a tof databases while the remaining ones had more restrictive access policies. We therefore fully support the recommendations published by other groups that governance models should be in place to facilitate data access, data sharing and secondary use of research data in health sciences.[37]

There are multiple challenges to the utilisation of EHDs in a regulatory context, particularly in Europe, which go beyond the above-mentioned challenges related to the characteristics of the specific databases. These include fragmentation and lack of interoperability of European data sources, inconsistent use of methods to integrate and analyse heterogeneous data, lack of systematic and

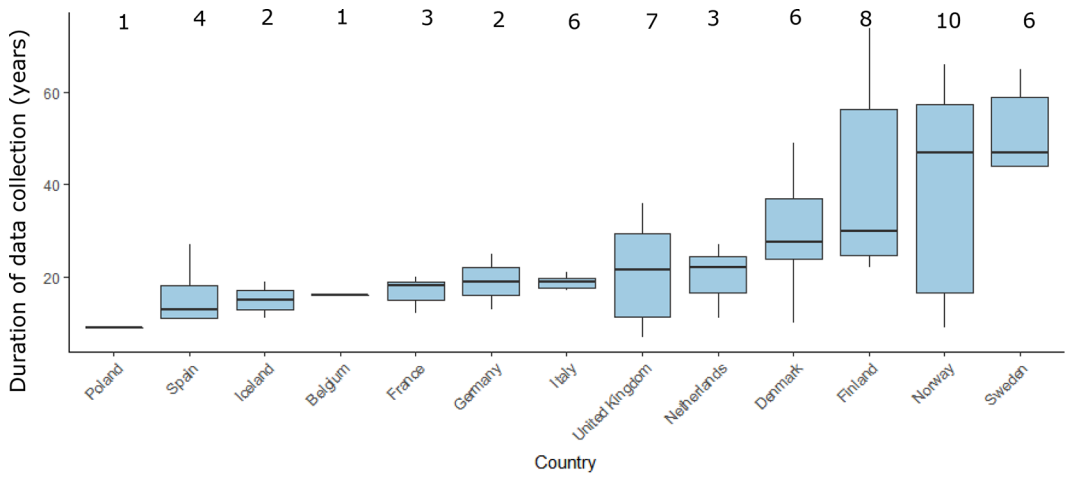

Figure 3 European data sources and duration of data collection. Box plots indicate the median (horizontal black line) data collection time by country while the margins of the box plot represent the IQR, the vertical lines indicate the minimum and maximum values. The number of databases per country are provided above the box plots.

consistent validation of data sources, governance issues and privacy concerns. In an attempt to deal with the significant heterogeneity across data sources in Europe, ENCePP has established a Working Group dedicated to facilitating the initiation and conduct of observational research using multiple data sources.[9] As part of its work, the group reviewed ongoing or finalised multidatabase drug safety projects of various publicly funded EU projects which highlighted the heterogeneity of the methods used for combining EHR data from multiple databases.[3] Ongoing work of the group is centred around developing guidance on conceptual models for multinational and multidatabase studies.

Our review has a number of limitations. First, we may have missed data sources during the identification process. However, we attempted to be as complete as possible by incorporating several rounds of database identification and review of the inventory by experts, including members of the ENCePP Working Group 'Data sources' and database owners. The difficulties we encountered when trying to map all the existing EHDs in Europe highlight again the need for more comprehensive and accessible repositories with EHDs.

Second, we excluded prescription-only databases since they cannot be used for aetiological studies even if we acknowledge their utility for drug-utilisation studies which are very common in the regulatory field. Lastly, validation of the primary source data is an important process that provides confidence in the results of the analyses,[38] and this was only evaluated indirectly through the number of validation studies reported by the database owners. A strength of our study was that data from publicly available sources was complemented or verified with database owners.

There is more work to be done in order to increase transparency and accessibility of existing datasources. Examples of areas for future development are to develop more robust validation measures and increase transparency of validated outcomes, to transform databases through a CDM to allow faster feasibility assessment and execution of studies, and to stimulate creation and access to EHRs in Eastern Europe.

CONCLUSION

We have provided a systematic inventory of EHDs available in Europe that includes a summary evaluation of their capability to support regulatory decision-making on the benefits and risks of medicines in Europe. Despite the wide range of healthcare databases available for epidemiological research in Europe, many of them were excluded from the inventory due to the absence of information needed for key regulatory activities. The analysis of the included databases confirmed the fragmentation, heterogeneity and lack of transparency existing in European EHDs.

The analysis has focused on population-based EHDs allowing conducting causal association studies between drug exposure and health outcomes in primary care. Our intention is to help the identification of and access to relevant existing databases that could be used for public health research. Beyond this objective, we consider that this inventory may assist clinical epidemiologists interested in undertaking other investigations such as studying the occurrence and determinants of health outcomes in a population.

We hope that this inventory should stimulate increased transparency and accessibility of other databases in addition to the development of data sources in Eastern European countries which are currently under-represented.

Contributors All authors (AP, KP,PMG, DRM, JS, DV, TG, XK and AC) were involved in the study design and data collection. AP, XK and DRM performed the analysis and interpretation of results. AP, DRM and AC contributed to writing and KP, PMG, JS, TG and XK revised and approved the final draft.

Funding This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Disclaimer The views expressed in this article are the personal views of the authors and may not be understood or quoted as being made on behalf of or reflecting the position of the European Medicines Agency or one of its committees or working parties.

Competing interests None declared.

Patient consent Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement An extended version of the dataset is available as supplementary material.

REFERENCES

1. Good pharmacovigilance practice. Module VIII. Post-authorisation safety studies. 2016 http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2012/06/WC500129137.pdf (accessed 17 Nov 2017).
2. Santoro A, Genov G, Spooner A, *et al*. Promoting and protecting public health: how the European Union Pharmacovigilance System Works. *Drug Saf* 2017;40:855–69.
3. Blake KV, Prilla S, Accadebled S, *et al*. European Medicines Agency review of post-authorisation studies with implications for the European Network of centres for pharmacoepidemiology and pharmacovigilance. *Pharmacoepidemiol Drug Saf* 2011;20:1021–9.
4. Mazzaglia G, Straus SMJ, Arlett P, *et al*. Study design and evaluation of risk minimization measures: a review of studies submitted to the european medicines agency for cardiovascular, endocrinology, and metabolic drugs. *Drug Saf* 2017.
5. Engel P, Almas MF, De Bruin ML, *et al*. Lessons learned on the design and the conduct of Post-Authorization Safety Studies: review of 3 years of PRAC oversight. *Br J Clin Pharmacol* 2017;83:884–93.
6. Carroll R, Ramagopalan SV, Cid-Ruzafa J, *et al*. An analysis of characteristics of post-authorisation studies registered on the ENCePP EU PAS Register. *F1000Res* 2017;6:1447.
7. Goedecke T, Morales DR, Pacurariu A, *et al*. Measuring the impact of medicines regulatory interventions - Systematic review and methodological considerations. *Br J Clin Pharmacol* 2018;84.
8. Sturkenboom M. *Other databases in Europe for the analytic evaluation of drug effects: pharmacovigilance*. : John Wiley & Sons, Ltd, 2007:73: 361.
9. Kurz X, Perez-Gutthann S. ENCePP Steering Group. Strengthening standards, transparency, and collaboration to support medicine evaluation: ten years of the European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP). *Pharmacoepidemiol Drug Saf* 2018;27:245–52.
10. Klungel OH, Kurz X, de Groot MC, *et al*. Multi-centre, multi-database studies with common protocols: lessons learnt from the IMI PROTECT project. *Pharmacoepidemiol Drug Saf* 2016;25(Suppl 1):156–65.
11. Trifirò G, Coloma PM, Rijnbeek PR, *et al*. Combining multiple healthcare databases for postmarketing drug and vaccine safety surveillance: why and how? *J Intern Med* 2014;275:551–61.
12. La Gamba F, Corrao G, Romio S, *et al*. Combining evidence from multiple electronic health care databases: performances of one-stage and two-stage meta-analysis in matched case-control studies. *Pharmacoepidemiol Drug Saf* 2017;26:1213–9.
13. Strom B. *How should one perform pharmacoepidemiology studies? Choosing among the available alternatives - pharmacoepidemiology*. Third edn: Wiley Online Library, 2013:401–13.
14. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J Clin Epidemiol* 2005;58:323–37.
15. van Staa TP, Klungel OH. Real-life data and learning from practice to advance innovation. http://www.who.int/medicines/areas/priority_medicines/BP8_4Data.pdf
16. Abbing-Karahagopian V, Kurz X, de Vries F, *et al*. Bridging differences in outcomes of pharmacoepidemiological studies: design and first results of the PROTECT project. *Curr Clin Pharmacol* 2014;9:130–8.
17. Coloma PM, Schuemie MJ, Trifirò G, *et al*. Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project. *Pharmacoepidemiol Drug Saf* 2011;20:1–11.
18. Madigan D, Ryan PB, Schuemie M, *et al*. Evaluating the impact of database heterogeneity on observational study results. *Am J Epidemiol* 2013;178:645–51.
19. Roberto G, Leal I, Sattar N, *et al*. Identifying cases of type 2 diabetes in heterogeneous data sources: strategy from the EMIF project. *PLoS One* 2016;11:e0160648.
20. Hall GC, Sauer B, Bourke A, *et al*. Guidelines for good database selection and use in pharmacoepidemiology research. *Pharmacoepidemiol Drug Saf* 2012;21:1–10.
21. ENCePP Resources Database. http://www.encepp.eu/encepp (accessed 1 Jan 2018).
22. Bridge to data. https://www.bridgetodata.org/ (accessed 1 Jan 2018).
23. Strom BL, ed. *Pharmacoepidemiology*. Chichester, UK: John Wiley & Sons, Ltd, 2000. ().
24. Rothman KJ. *Epidemiology: an introduction*. USA: Oxford University Press, 2002.
25. PARENT Joint Action. Methodological guidelines and recommendations for efficient and rational governance of patient registries. 2015 http://patientregistries.eu/deliverables (accessed 1 Jan 2018).
26. Boggon R, van Staa TP, Chapman M, *et al*. Cancer recording and mortality in the General Practice Research Database and linked cancer registries. *Pharmacoepidemiol Drug Saf* 2013;22:168–75.
27. Gagne JJ. Common models, different approaches. *Drug Saf* 2015;38:683–6.
28. García-Gil MdelM, Hermosilla E, Prieto-Alhambra D, *et al*. Construction and validation of a scoring system for the selection of high-quality data in a Spanish population primary care database (SIDIAP). *J Innov Health Inform* 2011;19:135–45.
29. ADVANCE Work package 5 white paper proof-of-concept studies. http://www.advancevaccines.eu/?page=publications&id=DELIVERABLES (accessed 23 May 2018).
30. Wettermark B, Zoëga H, Furu K, *et al*. The Nordic prescription databases as a resource for pharmacoepidemiological research--a literature review. *Pharmacoepidemiol Drug Saf* 2013;22:691–9.
31. OECD iLibrary. Health at a Glance: Europe 2016. OECD READ edition. https://www.oecd-ilibrary.org/social-issues-migration-health/health-at-a-glance_19991312 (accessed 25 Jan 2018).
32. EMA Annual Report. 2016 http://www.ema.europa.eu/docs/en_GB/document_library/Annual_report/2017/05/WC500227334.pdf (accessed 1 Jan 2018).
33. Neubert A, Sturkenboom MC, Murray ML, *et al*. Databases for pediatric medicine research in Europe--assessment and critical appraisal. *Pharmacoepidemiol Drug Saf* 2008;17:1155–67.
34. Gini R, Schuemie M, Brown J, *et al*. Data extraction and management in networks of observational health care databases for scientific research: a comparison of EU-ADR, OMOP, Mini-Sentinel and MATRICE Strategies. *EGEMS* 2016;4:2.
35. European Medical Information Framework. http://www.emif.eu/ (accessed 1 Jan 2018).
36. Research and Innovation. European Health Data Network. 2018 http://ec.europa.eu/research/participants/portal/desktop/en/opportunities/h2020/topics/imi2-2017-12-04.html (accessed 1 Jan 2018).
37. Burton PR, Banner N, Elliot MJ, *et al*. Policies and strategies to facilitate secondary use of research data in the health sciences. *Int J Epidemiol* 2017;46:1729–33.
38. Ehrenstein V, Petersen I, Smeeth L, *et al*. Helping everyone do better: a call for validation studies of routinely recorded health data. *Clin Epidemiol* 2016;8:49–51.